International Journal of Computer Science Engineering
and Information Technology Research (IJCSEITR)
ISSN (P): 2249-6831; ISSN (E): 2249-7943
Vol. 8, Issue 2, Jun 2018, 1-8

TRANS STELLAR
•Journal Publications •Research Consultancy

Original Article

ANALYSIS AND COMPARISON OF LOAN SANCTION PREDICTION MODEL USING PYTHON

SRISHTI SRIVASTAVA, AYUSH GARG, ARPIT SEHGAL & ASHOK KUMAR

Department of Information Technology, College of Technology, GBPUAT Pantnagar, Uttarakhand, India

ABSTRACT

With increase in loan seekers around the world while bank assets remain the same, it becomes utmost important for banks to find out the credible customers. In this paper, we have taken the previous records of people to whom the loan was granted and on the basis of those records, the system is trained using the machine learning model. The implementation has been done in four main sections i.e., Data Visualization, Data Munging, Building the Predictive Model and Testing.

The main objective that is fulfilled here is, by using this system, we will easily be able to identify the credible customers, which otherwise involves rigorous procedure by bank officials, in the present scenario.

KEYWORDS: Loan Sanction Prediction, Python, Logistic Regression, Decision Tree, Random Forest & Machine Learning

Received: Mar 07, 2018; **Accepted:** Mar 27, 2018; **Published:** Apr 23, 2018; **Paper Id.:** IJCSEITRJUN20181

INTRODUCTION

Sanctioning of loan to borrowers form the most vital part of every bank's business, as most of its assets come from the profit gained in the loan distribution process. Therefore, it is essential for banks to estimate whether the customer is right or not i.e., his ability to default or not in the coming future. By doing so, the bank would know that its assets are in safe hands. The process of choosing right customers can be automated by making use of machine learning algorithms. In this paper, we have designed a model that provides an easy, effective and serviceable way to choose deserving applicants.

The supervised machine learning models used in our proposed system are briefly described below:

- **Logistic Regression:** This model is based on classification technique, which predicts the outcome of a variable in the form of binary (0 and 1) or boolean (Yes and No) form. It can be considered as a special case of linear regression, when the outcome of the variable is sectional.
- **Decision Tree:** This is a classification model, based on tree like structure of decisions. It starts with a single node at the top, which branches out into possible outcomes. Each of these branches further leads to additional outcomes, making it a tree like structure or model.
- **Random Forest:** This classification model is a multitude of decision tree at the time of training, and gives a class as an outcome, which is the mode of the classes. It is better in a way that it removes the over fitting of decision tree on the data set.

To evaluate our predictive model, **K-fold cross validation** is applied. It is an approach to partition the dataset into training set, which is used to train the predictive model and test set which evaluates the error on the combination of different data sets. It arbitrarily divides the sample data into K equal size subsamples.

RELATED WORK

In [I], the author predicts the profitability of Default of a bank loan applicant, using the Data Mining functions of the R language. It checks the credibility of the applicant and saves the bank from approving loan to those people, who won't be able to repay it. The work in [II] predicts the credibility of the customer, to whom the loan has to be granted. Work proposed in [III] makes the use of Ensemble Modeling to improve the accuracy of the prediction. It shows how ensemble model compares several models and picks up the one, which provides the best result for the data set. Author in [IV] concluded that the Tree Model for Genetic Algorithm is the best amongst all the other models for forecasting the finance for customers. He did it, by experimenting five times on the data set using different models in the R language. Work in [V] uses the Decision tree algorithm to determine the approval/rejection of the loan request of the loan applicant.

PROPOSED MODEL

As the process of loan sanctioning is quite grueling and tedious, it involves a high risk of human error. The model proposed in this paper will analyze the data sets using the Random Forest algorithm. Algorithms like Logical Regression and Decision tree were also tested for the purpose, but their accuracy and cross validation (when combinedly analyzed over the top 3 parameters) turned out to be less than that of the Random Forest. The model mainly emphasizes on sanctioning the loan to the customer in the most accurate way possible, speeding up the process and minimizing the human work to a great extent.

The model is divided into three main sections

In this, we have taken the Kaggle datasets [VI] for Loan Prediction problem.

Data Exploration

This involves summarization and exploration of dataset. This is done by using python visualization libraries, matplotlib [VII] and plotly [VIII]. From this, we can get to know the distribution and relationship between different variables.

Data Munging

Data munging refers to the cleaning of dataset, in order to make it ready to feed it as input to data model. While exploration of datasets, we encountered the problem of non-integer data and some missing (null) values in the variables. For categorical data, encoding is done in order to change non-numerical values to numerical ones. The missing values are filled with Mean, Median or Mode as per the need of the variables. For complex variables that depend on many variables, we define certain functions in order to fill the null values.

Building Predictive Model

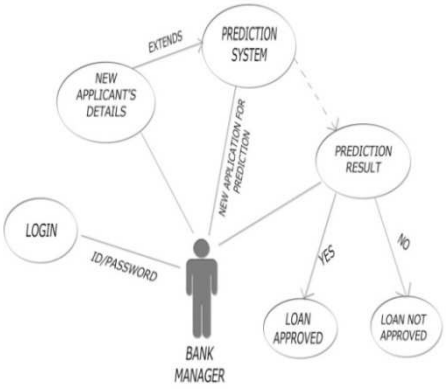
After we have made data useful for models, we made a predictive model on the dataset. For this, we used Scikit-Learn (sklearn) [IX] library. Generic classification function is defined, in which, the Accuracy and Cross-Validation scores are determined by taking model as an input. Different models are used to make prediction on new test cases.

USE CASE DIAGRAM AND THEIR TEMPLATES

Database Administrator

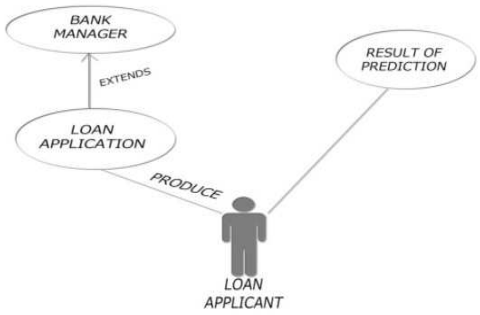
Basic Description	The basic requirement is the record of past loan applicants on which the predictive model is based.
Actors	Database Administrator
Flow of Control **1.Basic Flow :** **2.Alternate Flow :**	DBA enters the system by entering his login ID and password. None
Special Requirements	Should update the new applications in the database
Pre-Condition	Record of past applications
Post-Condition	Maintenance and good understanding of prediction system
Extension Points	None

Bank Manager



Basic Description	Bank Manager requires the details of new loan applicant in order to feed it into the system
Actors	Bank Manager
Flow of Control **1.Basic Flow :** **2.Alternate Flow:**	 Enters the system by his login ID/Password None
Special Requirements	All the parameters required for loan sanction prediction should be known
Pre-Condition	Knowledge about details of the new applicant
Post-Condition	Whether loan should be granted or not
Extension Points	None

Loan Applicant



Basic Description	The applicant needs to produce all the documents to the manager with his/her complete details
Actors	Loan Applicant
Flow of Control **1.Basic Flow :** **2.Alternate Flow:**	 Gives his/her details which is then fed to system by the manager None
Special Requirements	None
Pre-Condition	None
Post-Condition	Knows the result whether loan would be sanctioned or not
Extension Points	None

ALGORITHM

Start

Step 1: Extract the given data.

Step 2: Replace all null values (Data Munging)

Step 3: Encode (Non-numeric Data = Numeric Data)

Step 4: Define function classification_model()

Step 4.1: Calculate accuracy

Step 4.2: Calculate error using K-Fold Cross Validation

Step 4.3: Calculate Cross Validation score.

Step 5: Obtain weightage of each variable in descending order.

Step 6: Set parameters for n variables and call function in STEP 4.

Step 7: Compare results of all model.

Step 8: Fetch new data for prediction.

Step 9: Return or Print the result.

End

WORKING OF ALGORITHM

- Extract the data from csv file into a data frame using pandas [X] library.
- Replace any null value present in the data with suitable value calculated after proper analysis of the data (data munging).
- Encode any non-numeric data present to numeric values.
- Define a "classification model" function to make a classification model and for accessing performance.
- Accuracy is calculated on the predicted values.
- Error is calculated using the K-fold cross validation with 5 folds.
- Cross validation score is calculated by taking the mean of all the errors.
- Obtain the weight age of each variable in descending order of their weight.
- Take each model, and calculate accuracy and cross validation score for both top 3 variables and top 5 variables of the weight age list, using the "classification model" function in step 4.
- Compare the results of all models, and choose the most efficient model with high accuracy and cross validation score.
- Use this model along with its respective variables to predict the outcome of any new data set entered by a user through the webpage.
- Return or print the final status to the user.

IMPLEMENTATION

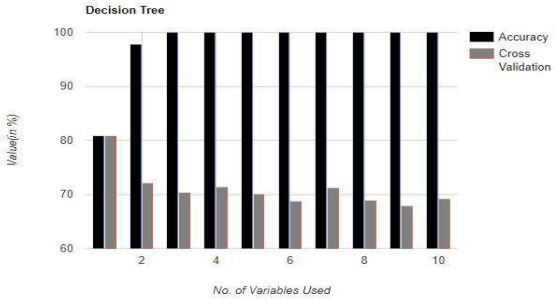
Weightage of each Variable at an Instance (Decision Tree)

S. No.	Variable Name	Weight age
1	Total Income_log	0.298756
2	Credit_History	0.292201
3	Loan Amount_log	0.233202
4	Dependents	0.040203
5	Property_Area	0.032426
6	Education	0.032147

Table: Contd.,		
7	Loan_Amount_Term	0.029820
8	Self_Employed	0.019573
9	Gender	0.013639
10	Married	0.008033

Decision Tree

The number of variables is taken from top to bottom in decreasing order of the weightage of variables, as shown above for Decision Tree.



No.	Accuracy	Cross-Validation
I	95.9283	56.8439
II	97.8827	72.1551
III	100	70.3585
IV	100	71.3368
V	100	70.0293
VI	100	68.7285
VII	100	71.1715
VIII	100	68.8884
IX	100	67.9142
X	100	69.219

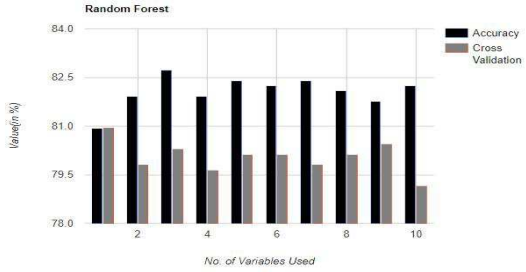
We can observe that in decision tree model, the accuracy is high, but the cross validation score is quite less. This means that the model is over fitting the training data and may not be able to accurately predict on new data sets.

Weight Age of each Variable at an Instance (Random Forest)

S.No.	Variable Name	Weightage
1	Credit History	0.273205
2	TotalIncome_log	0.264097
3	LoanAmount_log	0.224286
4	Dependents	0.054631
5	Property_Area	0.049464
6	Loan_Amount_Term	0.044424
7	Married	0.026957
8	Education	0.023202
9	Gender	0.020303
10	Self_Employed	0.019431

Random Forest

The number of variables is taken from top to bottom in decreasing order of the weightage of variables as shown above for Random Forest.

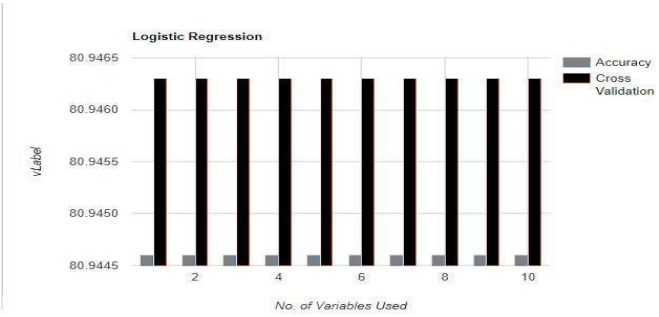


No.	Accuracy	Cross-Validation
I	80.9446	80.9463
II	81.9218	79.8081
III	82.7362	80.2972
IV	81.9218	79.6415
V	82.4104	80.1346
VI	82.2476	80.1306
VII	82.4104	79.8081
VIII	82.0847	80.1319
IX	81.759	80.4598
X	82.2476	79.159

We can observe that in random forest, the accuracy and cross validation score is quite good with 3 variables, as compared to other combination of variables.

Logistic Regression

The number of variables is taken from top to bottom in decreasing order of the weightage of variables, as in Random Forest.



No.	Accuracy	Cross-Validation
I	80.9446	80.9463
II	80.9446	80.9463
III	80.9446	80.9463
IV	80.9446	80.9463
V	80.9446	80.9463
VI	80.9446	80.9463
VII	80.9446	80.9463
VIII	80.9446	80.9463
IX	80.9446	80.9463
X	80.9446	80.9463

We can observe that the accuracy and cross validation are same, irrespective of the number of variables used.

Therefore, this model is not suitable for the prediction model.

CONCLUSIONS

In the proposed work of Loan Sanction Prediction, three algorithms - Logical Regression, Decision Tree and Random Forest were used to calculate the accuracy of the prediction. By continuous test on the datasets, we inferred that the Accuracy and Cross-validation score of Random Forest is the highest amongst all the three models. Therefore, the prediction is done using the Random Forest Algorithm. Main focus of this application is to help the bankers, to sanction the loan to valid applicants in a short span of time. By the proper examination of the algorithms, we can conclude that the software is very efficient, and will help the bankers to sanction the loan much efficiently.

REFERENCES

1. *Sudhamathy G (2016). Credit Risk Analysis and Prediction Modelling of Bank Loans Using R. International Journal of Engineering and Technology (IJET)*
2. *Anchal Goyal, Ranpreet Kaur (2016). Loan Prediction Using Ensemble Technique. International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE)*
3. *Anchal Goyal, Ranpreet Kaur (2016). A survey on Ensemble Model for Loan Prediction. International Journal of Engineering Trends and Applications (IJETA)*
4. *KAUR, SATWANT, RISHMA CHAWLA, and VARINDERJIT KAUR. "Implementation and Evaluation of Optimal Algorithms Based on Decision Tree & Clustering Algorithms."*
5. *Anchal Goyal, Ranpreet Kaur (2016). Accuracy Prediction for Loan Risk Using Machine Learning Models. International Journal of Computer Science Trends and Technology (IJCST)*
6. *Sivashree MS, Rekha Sunny T (2015). Loan Credibility Prediction System Based on Decision Tree Algorithm. International Journal of Engineering Research & Technology (IJERT)*
7. *https://datahack.analyticsvidhya.com/contest/practice-problem-loan-prediction-iii/*
8. *https://matplotlib.org/*
9. *https://plot.ly/python/*
10. *http://scikit-learn.org/stable/_downloads/scikit-learn-docs.pdf*
11. *https://pandas.pydata.org/pandas-docs/stable/pandas.pdf*